# A SUGGESTED USE OF SEQUENTIAL ANALYSIS IN PERFORMANCE ACCEPTANCE TESTING

Prepared for
Personnel Analysis Division
Bureau of Naval Personnel
Contract N6ori - 07142
Project NR153-124

by
Rupert N. Evans, Ph.D.
College of Education
University of Illinois
Urbana, Illinois

TABLE OF CONTENTS

## I. INTRODUCTION

One of the chief obstacles to the more widespread use of performance testing has been the relatively greater time and expense required by performance tests as compared with paper and pencil tests. Many performance tests are individually administered, and require the use of a highly trained observer and a piece of expensive equipment for one or two days in order to administer ten or twenty items to one individual. During one or two hours, one test administrator can give perhaps fifty to two hundred paper and pencil items to a large group of subjects in an ordinary class room. Obviously on a time and expense basis, paper and pencil tests are much more acceptable. That performance tests have continued to be used at all is a testimony to strong feelings about their usefulness.

Performance tests are ordinarily used for one or more of three purposes: (1) to determine whether or not the person tested possesses certain qualities to a desired degree--acceptance testing; (2) to determine in what areas the person tested needs further training--remedial testing; (3) to enable the person tested to perform certain activities more skillfully--instructional testing.

The first of these purposes--the use of performance tests for acceptance--is most widely used. Typical situations calling for this sort of testing are for the determination of: (1) graduation from a particular course or phase of a course; (2) acceptability for skilled employment; and (3) advancement in rating. Anytime you desire to know whether a particular applicant passes or fails, is successful or unsuccessful, is desirable or undesirable, you can use performance tests as the basis for acceptance or rejection.

The second and third uses of performance testing, remedial and instructional, are seldom used outside of training programs. While they are important, this discussion is not concerned with them primarily.

Regardless of the purpose or purposes for which they have been designed, performance tests have been administered, ordinarily, as a block. That is, each person tested is given the same number of performance test items. Paper and pencil tests also have been administered as a rule in this same manner, with every person taking the test being given the same number of items. It is the purpose of this discussion to show that this is not necessarily the most economical procedure, particularly for performance acceptance testing. Sequential sampling is proposed as an alternative. Since sequential sampling has been used most widely in industrial acceptance of supplies, let us take a look at sequential analysis as it is used in industry at the present time, to see if it has application to performance testing of people.

## Acceptance Testing of Supplies in Industry and the Armed Forces

It is usually necessary for any agency purchasing commodities from some other agency to determine the quality of the products which are supplied to it. If you set out to buy a quantity of receivers from some vendor, you establish specifications for these receivers, and then check the receivers which are shipped to you to see whether or not they are acceptable. Similarly, if one section of your organization produces hydraulic fittings which will later be assembled into a gun mount, you ordinarily will check the fittings to see whether they are acceptable before shipping them to final assembly.

One method of determining acceptability of products is to give them a one hundred per cent check. This obviously is impractical if the test destroys the product being tested. But even with non-destructive tests, most industrial concerns have adopted sampling procedures for determining the acceptability of a lot (group of products). Usually the procedure has been to determine that a certain number or a certain percentage of a lot would be checked, and the acceptability of the whole lot determined from the sample. (This is basically the same procedure we use in testing people. If we want to determine a person's grade in a course at Annapolis, we pick a sample of the almost infinite number of questions we could ask about the course, and estimate the percentage of questions he <u>could</u> answer from the percentage of questions he <u>did</u> answer correctly on the test.) Note that with this procedure, the number of items to be tested is determined before testing is begun.

Since World War II, a method of sampling called "sequential sampling" has been coming into wide use in quality control in industry and the armed forces. Essentially, this method of sampling requires that a small sample of the lot be tested. Then on the basis of this sample, one of three decisions is reached: (1) accept the lot, (2) reject the lot, or (3) continue testing. If it is decided to continue testing, another sample is inspected, and on the basis of this information plus the information from the preceding lots, one of the above three decisions is made. Sampling is continued until the lot can be accepted or rejected. This method requires much smaller samples for the lots that are extremely good, or extremely poor. For the few lots that are on the borderline between acceptance or rejection, you may sample as many or more items than are required when every lot is tested in exactly the same way. Almost invariably, however, for a given degree of confidence in the results, sequential sampling will require fewer tests than are required by conventional sampling. (Wald, in his book, <u>Sequential Analysis</u>, estimates the average saving at about 50 per cent.)

There appear to be good reasons why sequential sampling, which has proved so successful in acceptance sampling in industry and the armed forces, can be applied with success in acceptance testing of people.

## Probable Limitations of Sequential Sampling in Testing People

One feature of sequential sampling makes it difficult to employ with certain types of tests when you are testing people instead of commodities. Before one can proceed from the first to the second test item in a group of test items, it is necessary to know the score a person has made on the first item, or at least whether the person has passed or failed that item. For group tests of the paper and pencil type, a person could be taking several items during the time required for the first one to be scored. Thus, the chief advantage of sequential analysis, a saving of time and expense, is lost. For individual performance tests, however, the time required to determine a person's rating on a test item is insignificant in comparison with the time required for giving additional, possibly unneeded items.

Sequential sampling is not as good for diagnostic or instructional testing as are conventional tests, because for diagnosis and instruction it is commonly desirable to expose the student to a wide range of items, rather than to conclude the testing in as short a time as possible. This caution does not necessarily apply to acceptance testing, however.

For best use with sequential sampling, the separate items on a test need to be as nearly alike as possible, that is, there should be high item intercorrelation. This is desirable in order to increase the confidence you can place in the results of almost any test, but it is particularly important with sequential analysis. In order to maximize item intercorrelation, it appears most desirable to use sequential sampling to determine whether a person passes or failsa phase of a course, since here the items will be very much alike. A final examination for determining whether a person passes or fails a long course would not be as good, because items would cover a wide breadth of material, and would be much less alike. The use of sequential sampling for determining whether a person passed or failed a practical factor examination for advancement in rating would probably be better than an examination over a long course, but poorer than an examination over one phase of a course.

## Advantages of Sequential Sampling in Testing People

When the above limitations are recognized, and necessary precautions are observed, sequential sampling offers one tremendous advantage in testing people. Those persons who are extremely poor or extremely good can be rejected or accepted after a relatively short period of testing.

Whether we recognize it or not, whenever we set up a test, we set certain confidence limits in the results. Ordinarily, the greater the length of the test, the greater confidence one can place in the results. When you have decided what level of confidence you wish, by using sequential sampling you can test extremely poor or extremely good persons with far fewer items than are necessary for those people who are near the cutting point in "true" ability. In fact, in certain situations, one or two items

are as reliable for those people at the extremes of ability as ten or twenty items are for the person near the cutting point. Consequently, when the cost of testing is high, as with performance tests requiring one-half to two hours per test item, a marked saving can be made by employing sequential sampling at no sacrifice of the overall reliability standard.

In plain language, when you give the same number of items to people of varied ability, you can place much more confidence in your test results for those people at the extremes of ability than you can for those people who are near the borderline between acceptance and rejection. When you use sequential sampling, you employ fewer test items for those people at the extremes than for those near the cutting point, and you place approximately equal confidence in the results for each level of ability.

## Summary

Performance tests as ordinarily administered require a great deal of time and expense. Sequential sampling, as adapted from industrial quality control, offers good possibilities of reducing this time and expense for acceptance testing. Sequential sampling involves taking a sample of performance and then deciding whether to (1) accept the lot or person, (2) reject the lot or person, or (3) continue testing.

## II. SAMPLING IN TESTING

When the ability of people is being tested, it is often a useful assumption than an infinite number of test items are available for testing a certain trait. For example, if you want to test the proficiency of an electronics technician on the practical factors involved in his rate, there is an almost unlimited number of performance items you can give to him. This number of items is so large that it can be considered to be nearly infinite.

Naturally, in any practical test we cannot give all of the test items which it is theoretically possible to give. Instead, we have to give a test made up of a sample of those theoretically possible. There are a number of reasons for this. Some items may involve too much expense, some are too dangerous to personnel or equipment, some cannot be scored consistently, etc. Even after we have eliminated all of the items which are impractical to administer, another practical consideration forces us to use only a sample of the remainder: only a certain amount of time can be made available for testing. Thus it is safe to assume that any test which is administered to people is a test which involves only a relatively small sample of the possible items.

Now we are not really interested in a man's ability to answer or do sample tasks. We want to know his true ability; that is we want to know how he would perform on the total number of possible items. However, we cannot get at a man's "true" ability except by using the sample items as a measure of his "true" ability.

Other things being equal, the larger the sample tested, the better the picture we get of the person's true ability. It is not at all uncommon to find one hundred or more items included in one paper and pencil test. So many items are used in an effort to boost the reliability of the test to get a more accurate idea of the man's true ability.

In most performance testing, however, it is impractical to give more than ten or twenty test items because of the time required per item, and because the test must usually be administered on an individual, rather than a group basis. The chief reason performance tests continue to be used is because people ordinarily feel that they measure very important aspects of a person's job that cannot be tapped by paper and pencil tests. Obviously, though, a performance test must involve only a sampling just as any other test. And if this sample does not give a good picture of a person's "true" performance, it is worthless. Thus we must make a compromise between a very long performance test with many items in the sample and a short performance test which will not interfere with other needs of the testing agency.

Fortunately, if our need is for acceptance testing, the problem is not quite so acute. In acceptance testing we do not need to get a complete picture of each man's "true" ability. We need only to determine whether he is above or below some standard. That is, does he or does he not make a passing mark. However, we do need to make sure, with a reasonable degree of confidence, that those persons whose true ability is above the cutting point are passed, and those whose true ability is below the cutting point are failed. Consequently we must still be concerned with test reliability.

For many purposes it is more convenient to think of a concept labelled "probability of acceptance" than to think of "reliability" when we are discussing acceptance testing. Probability of acceptance can be abbreviated $P_a$. One way of looking at probability of acceptance is to use the so-called "operating characteristics curve."

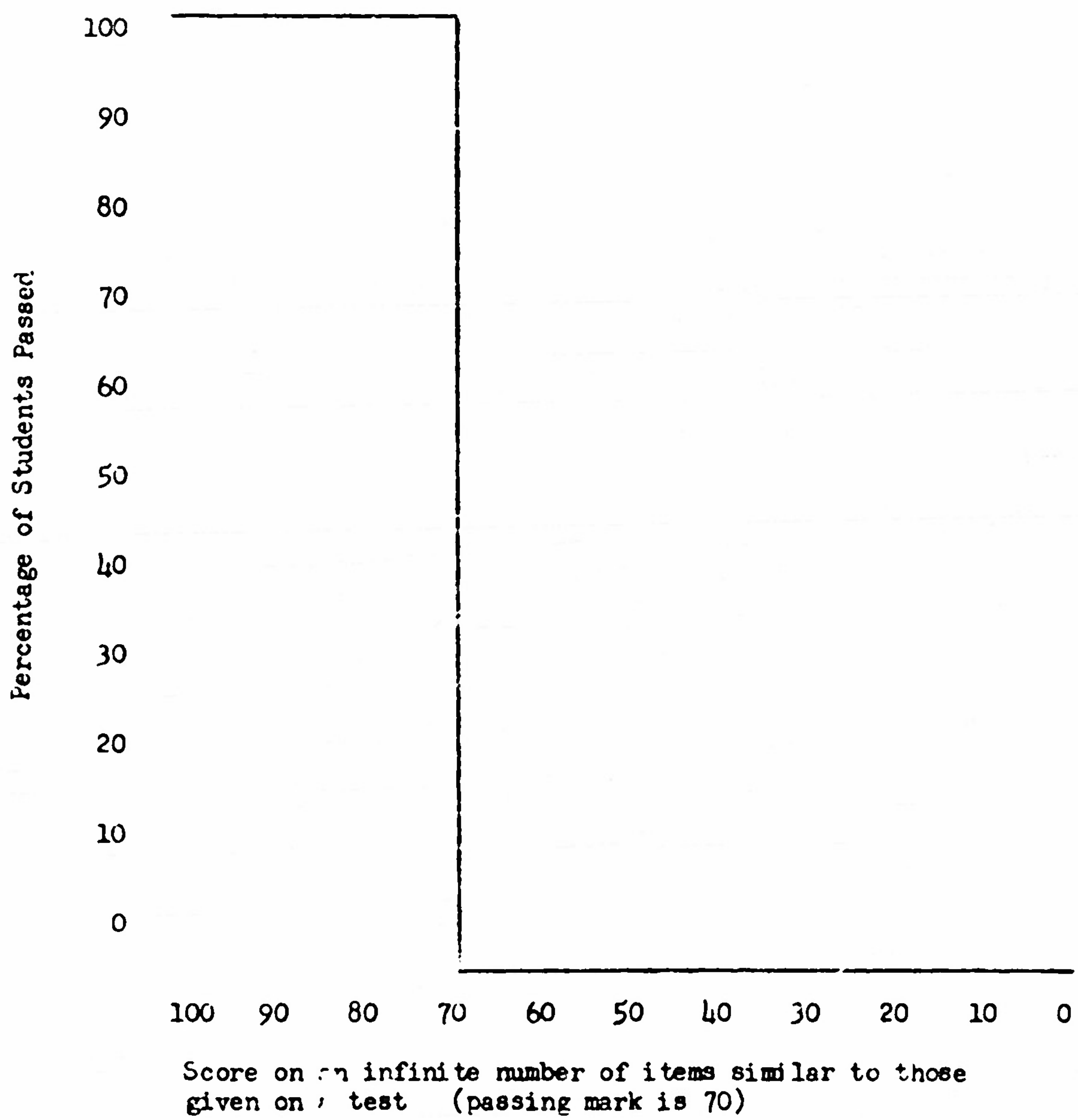
## The Operating Characteristics Curve

Ideally, when we set a "cutting score" (sometimes called a "passing mark" or "borderline between acceptance and rejection"), we want an operating characteristics curve like that shown in Figure 1.

This ideal curve would have a vertical line immediately above the borderline between acceptance and rejection. If it were possible to get such an ideal curve, and the "cutting score" or "passing mark" were set at 70, one hundred per cent of those who could make a score of 70 or better on an infinite number of such items would be passed (accepted), while all those who could make a score of less than 70 on an infinite number of such items would be failed (rejected). Note that the horizontal axis refers to the "true" score an individual would make on an infinite number of items, not to the score he would make on a practical test. Figure 1 is a theoretical curve probably never encountered in practice. Yet it is the sort of curve we want to strive for in acceptance testing. (Incidentally, a perfectly reliable acceptance test would show a curve like that in Figure 1. And we should remember that reliability places an upper limit on validity.)

FIGURE 1

Ideal Operating Characteristics Curve for Acceptance Testing (Ideal O C Curve)



Perhaps it will help us to get a better picture of the operating characteristics curve if we look at some OC curves for conventional tests of a length commonly used in performance testing. Figures 2, 3, and 4 show a series of operating characteristic curves for five-item tests

Poisson Approximation of Operating Characteristic Curves for Fixed Length (Conventional) Tests, (adapted from Grant, Statistical Quality Control, page 323 and Table G)

Vertical axis = % of students passed
Horizontal axis = True score, in % of items correct

FIGURE 2

Fig. 2 - Five items in test, Passing grade, 60%

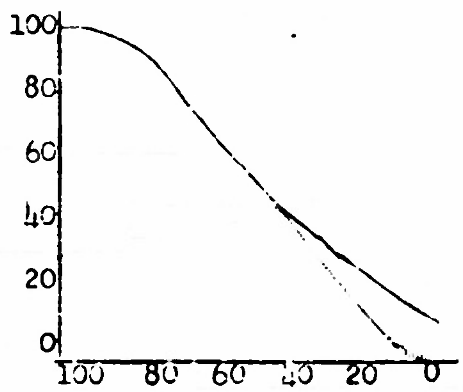



FIGURE 3

Fig. 3 - Five items in test, Passing grade, 80%

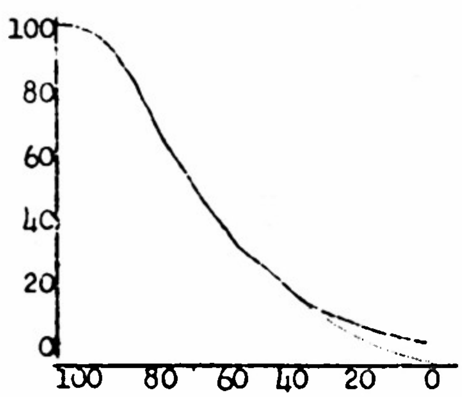
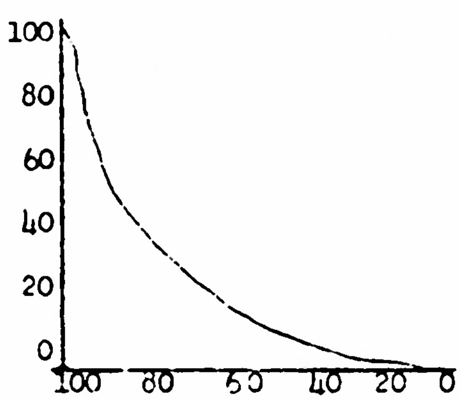



FIGURE 4

Fig. 4 - Five items in test, Passing grade, 100%



with passing marks of 60%, 80% and 100% respectively. Each person tested has been given all five of the items. Each of the items on these tests has been scored as either pass or fail. Note that in Figure 2, if you had a group of people whose "true" score was 70%, only 80% of these people would be accepted (pass the test). Moreover, if you had a group of people whose "true" score was only 40%, approximately 40% of them would be passed, even though their true score was far below the cutting score set for the test. Figures 5, 6, and 7 present similar information for a series of twenty item tests with passing marks of 60%, 80%, and 100%. Each person tested has been given all of the twenty items, and each of the items has been scored on a pass-fail basis. Here the picture is somewhat better, but it is still a long way from the ideal curve shown in Figure 1.

Vertical axis = % of students passed
Horizontal axis = True score, in % of items correct

FIGURE 5

Fig. 5 - Twenty items in test, Passing grade, 60%

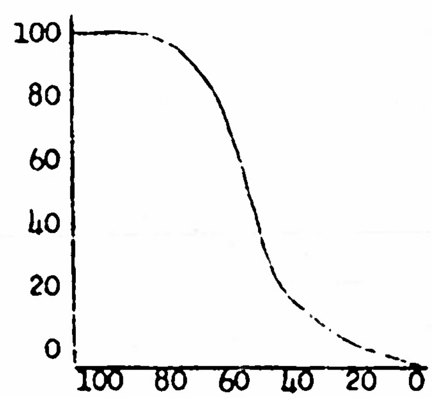



FIGURE 6

Fig. 6 - Twenty items in test, Passing grade, 80%

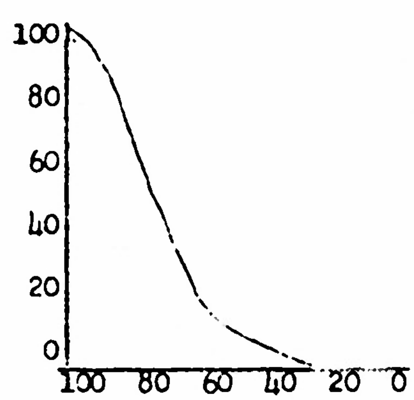
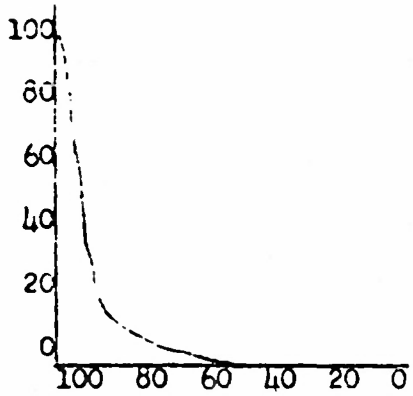


FIGURE 7

Fig. 7 - Twenty items in test, Passing grade, 100%




Other things being equal, the larger the number of items on the test, the closer the OC curve will approach Figure 1. The reason for this is that with a small number of items, there is a good chance that several of the items on the test happen to be among the few that a person of poor ability knows. Conversely, if there is a small number of items there is a good chance that several of the items on the test happen to be among the few that a person of great ability happens not to know. With a large number of items, this chance factor becomes less important.

The concept of the operating characteristics curve is one of the most important in acceptance testing. Without it, one is apt to fall into the common error of accepting test results as being necessarily a true picture of a man's ability.

## Summary

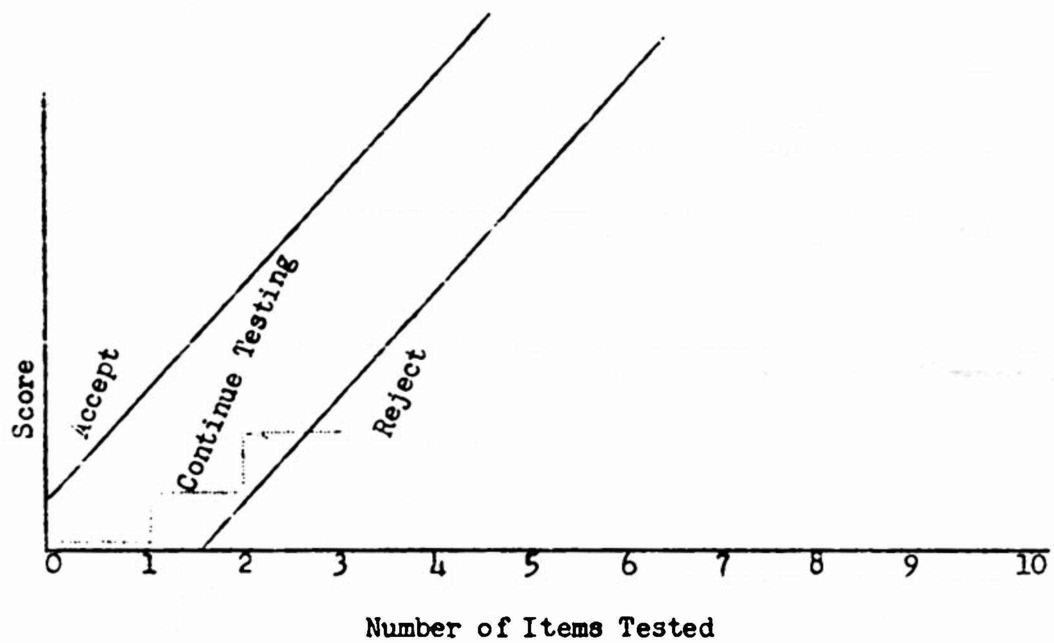
Any test should be regarded as a sampling of a great number of possible test items, and consequently test results are not necessarily a good picture of a man's true ability. The operating characteristics curve helps to show this discrepancy in acceptance testing. Other things being equal, the longer the test, the more reliable the results. Performance tests, however, are by nature limited in length, and some compromise must be reached with reliability.

## III. CHOOSING A SEQUENTIAL SAMPLING PLAN

All sequential sampling plans which are now used in industry are alike in that they can be represented by the general type of chart shown in Figure 8. As each item is given, the person's cumulative score is plotted above the item number. Testing continues until the graph runs

FIGURE 8

Graphic Presentation of a Sequential Sampling Plan




outside the two parallel lines into the "accept" area or the "reject" area. That ends the test. In the example shown in Figure 8, a small score was made on item number one; approximately the same score was earned on item two, but on the third item a score of zero was received, and the test ended with the rejection of the person tested. The slope and origins of the two lines determine how rapidly this will occur. (Because of the fact that most performance test items differ in difficulty and in discriminative value, a modification of Figure 8 is proposed for use in testing people. This modification is described in the next chapter.)

## Information Needed for Choosing a Sequential Sampling Plan

Two sorts of information are needed before a sequential sampling plan can be chosen for testing people: (1) what constitutes an acceptable and an unacceptable person; and (2) what risks are you willing to take of accepting a "poor" person and of rejecting a "good" person.

## Acceptable and Unacceptable Persons *

Ordinarily when a test is given for purposes of accepting or rejecting individuals in a group, some sort of passing score is used. All of those who score above this cutting point are passed, and all of those below are failed. However, when sequential analysis is used, two points are used, rather than a single cutting point.

One of these points may be described as the lower limit, or lowest score characteristic of the really good people. This point may be designated as $m_g$.

The other point may be described as the higher limit or highest score characteristic of the really poor people. This point may be designated as $m_p$. For example, the lower limit of really good people may be set as a score of 70 on a particular test, while the upper limit of really poor people is set as 50. In this case $m_g = 70$, and $m_p = 50$. The scores between $m_g$ and $m_p$ form the "zone of indifference" and are characteristic of people who are neither really good nor really poor.

This determination of the lower limit for really acceptable people, and the upper limit for really unacceptable people is the first decision that must be made in choosing a particular sequential sampling plan.

## Probability of Acceptance

Any sampling plan, whether it is a traditional test, a common performance test, an industrial inspection scheme, or a sequential test, involves certain risks. These risks are primarily due to sampling errors as discussed in the previous chapter, plus errors due to the instability of a man's performance from time to time. The second decision that must be made in setting up a sequential sampling plan involves the determination of the risk that you are willing to take for each of the two points discussed in the preceding paragraphs.

---

* The assumption is made throughout this discussion that all people at one end of a scale of ability are acceptable and all those at the other end of the same scale are unacceptable. This is the usual case in performance testing. If, however, you wished to accept only those people who were not too high or not too low on a scale (such as finger dexterity), much of the following discussion would not apply.

Naturally you want to be certain of accepting practically all of the "good" people, and certain of rejecting practically all of the "poor" people. However, the more certain you are, the greater the number of test items required. In the example above, if you are ready to take the risk of rejecting 5 out of 100 people whose true score ($m_g$) equals 70, the probability of acceptance of $m_g$ would equal .95. In other words you would be willing to buy a plan which in the long run would guarantee your acceptance of 95% of the people whose true score was 70. (This would be abbreviated $Pa_g = .95$, $m_g = 70$, which may be interpreted as: the probability of acceptance equals .95 when the true score equals 70.)

Similarly, if you are willing to take the risk of accepting 20 out of 100 people whose true score ($m_p$) equals 50, the probability of acceptance of $m_p$ would equal .20. In other words, you would be willing to buy a plan which in the long run would guarantee your acceptance of only 20% of the people whose true score was 50. (This would be abbreviated $Pa_p = .20$; $m_p = 50$.)

The closer $Pa_p$ is to zero, or the closer $Pa_g$ is to one, the more test items you will need to administer. That is, the more certain you want to be in your judgments, the more it will cost you.

Note that in the example above, $Pa_g$ was closer to one that $Pa_p$ was to zero. We wanted to be more certain of getting all of the "good" persons than we wanted to be certain of rejecting all of the "poor" persons. This is the usual situation when many men are needed, particularly when there are going to be other opportunities later on of weeding out the poor people who were inadvertently accepted. However, there is nothing to prevent the risk of accepting $m_g$ from equalling the risk of rejecting $m_p$ (for example, $Pa_g = .90$, and $Pa_p = .10$). Or when there is an over supply of men, or when the acceptance of a poor man may mean serious consequences such as the failure of a mission, the risk of accepting poor men may be set lower than the risk of rejecting good men (for example, $Pa_g = .80$, and $Pa_p = .001$). During World War II, the Office of Strategic Services had many men to choose from and vital missions to perform, so they were willing to take the chance of rejecting many good men, provided that they could be reasonably certain of getting very few poor ones.

The choice of $m_g$, $m_p$, $Pa_g$, and $Pa_p$ determines the operating characteristics curve of the sequential sampling plan.

## Means of Reporting Scores

Performance test data are usually available in a variety of forms. With a simple method of scoring, results may be reported as "pass" or "fail". More precise scores are usually expressed numerically, with as many as one hundred or more different grades possible for one test item. Sometimes letter grades are used in reporting scores.

The method of reporting scores, whether numerical, letter, or word grades, is relatively unimportant. It is important, however, to consider the number of scores which students actually can make on a test item. Other things being equal, it is possible to learn much more about a student's performance if he can make any one of five possible scores on an item, than if his performance is reported as either "pass" or "fail". As one would expect, for a given $m_g$, $m_p$, $Pa_g$, and $Pa_p$, it ordinarily requires far fewer test items to determine acceptance when five or ten scores are given on each item than when only two scores are available.

There is ordinarily a practical limit to the number of test scores which should be attainable on any one item. If more than about ten scores are reported, the calculations necessary for sequential analysis become rather laborious. If, for example, time in seconds required to perform some task were used as a grade, those scores which students actually attain can be grouped to bring the total number of scores within a reasonable limit.

As is usually the case, however, other things are not always equal. Sometimes you are much surer of your judgments in evaluating a performance item if you report it as passed or failed, rather than in terms of a score. Many times it is much quicker to score an item as passed or failed. A common example is in the use of objective and essay type paper and pencil examinations. Objective questions are almost invariably scored as either right or wrong. Essay questions could be graded with a score, or as pass-fail, but are usually given a score. Yet objective questions are widely used because they are easier to score and because they can be administered more rapidly. It is recommended as a general rule that five or more scores per item be used in sequential analysis whenever practical, and that when more than ten scores are used, scores be grouped to provide somewhere between five and ten intervals.

## Item Intercorrelation and Difficulty

The original use of sequential sampling was in acceptance inspection for the armed forces. In this type of sampling the problem is to determine whether a lot is acceptable by testing a sample from that lot. This involves making the same test on each of the products in the lot. When you determine a person's acceptability as a trouble shooter on radar gear, the problem is to determine his acceptability by testing him on a sample of radar trouble shooting problems. The total range of that person's trouble shooting ability is comparable to the industrial lot; the group of trouble shooting items you give to that person is comparable to the inspection sample drawn from the industrial lot; and one test item for that person is comparable to a test of one piece from the industrial inspection sample.

This analogy breaks down somewhat on the last comparison. Each piece in the industrial inspection sample is given exactly the same test. In most cases we cannot give one person a series of test items exactly alike, for if he knew the answer to one of them, he would know the answer to all. We would really be giving him only one test item a number of times. (This would be all right if we were testing a basketball player's ability to shoot free throws, but where any sort of problem solving or progressive learning is important in the test, it is impractical to repeat identical test items.)

In practice if we want to measure trouble shooting ability of a certain type, we prepare several different trouble shooting test items which are very similar. All of them would involve trouble shooting on a particular type of radar gear, for example. If, through this procedure, you obtain relatively high item intercorrelation, a large part of your problem is solved. However, it is also necessary that item intercorrelations be insignificant when the criterion score (internal or external) is held constant.

One further step is necessary in order to insure equal difficulty level. In the industrial inspection situation, each test on each commodity in the sample is of equal difficulty, since each test is the same. This is not true of the usual type of item which must be chosen for testing people. Even though you obtain high item intercorrelation, some test items will be much easier than others. It is particularly important that this be taken into account in sequential sampling, for if several very easy items were chosen by chance to be the first items administered, and no adjustment were made in scoring, practically everyone tested would be accepted right away.

One possible solution for the problem of item difficulty is to convert test scores into some standard score such as the "T" score. An even simpler method of compensating for item difficulty empirically is described in the next chapter.

## The Average Sample Size Curve

In sequential analysis, unlike most methods of testing, there is no way to know in advance how many test items are required for determining the acceptability of any one person. * It is possible, however, to determine the average sample required. The curve showing this information is known as an "Average Sample Size Curve".

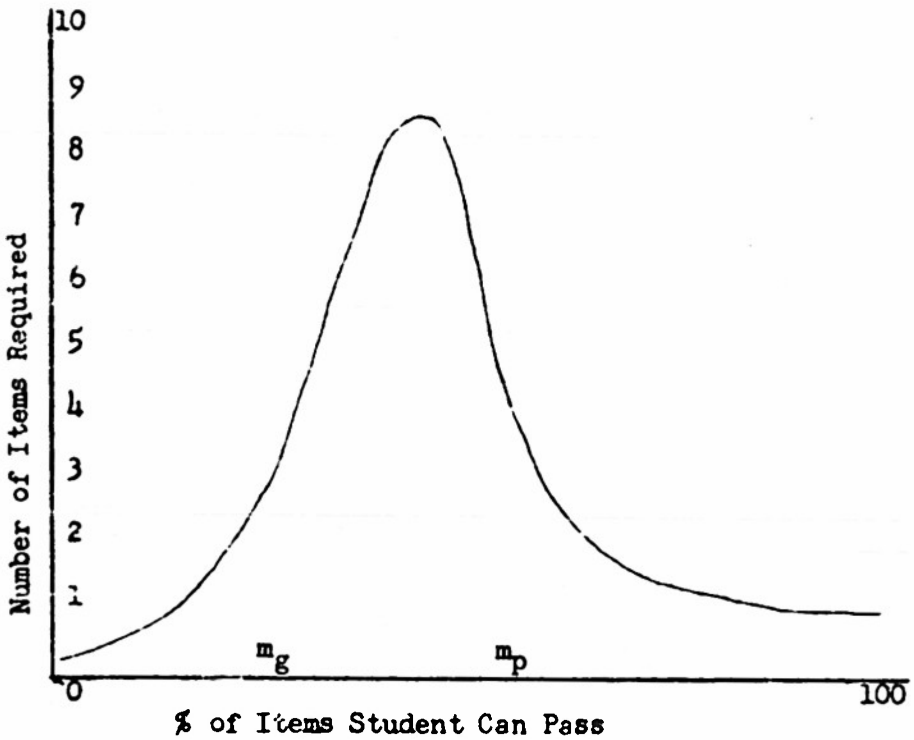
---

* In practice a decision is usually made to stop testing after a certain number of items, and to declare the person tested to be accepted (or alternatively, to stop testing after a certain number of items, and to declare the person tested to be rejected).

The Average Sample Size curve has a characteristic shape. (See Figure 9.) The largest number of items is required in testing persons

FIGURE 9

Typical Shape of Average Sample Size Curves




who are not "good" or not "poor" (those who are between $m_g$ and $m_p$ in ability). The height of the curve is determined by $Pa_g$ and $Pa_p$, the probability of accepting "good" and "poor" persons, respectively. Its position over the base line is determined by $m_g$ and $m_p$. The procedure for computing the average number of test items required for a particular sequential sampling plan is presented in the next chapter.

## Summary

The choice of a sequential sampling plan is determined by decisions as to (1) what constitutes an acceptable and an unacceptable person; and (2) what risks can be taken of accepting a "poor" person and of rejecting a "good" person. Ordinarily, a person should be able to make one of five or more scores on any one test item. The number of items required for testing a particular person can never be known in advance, but the average number of items required can be determined.

## IV. TENTATIVE SUGGESTIONS FOR PUTTING A SEQUENTIAL SAMPLING PLAN INTO OPERATION *

It is assumed that a performance test has been constructed according to accepted principles such as those outlined in standard reference works in this field.[1] A few more test items should be planned than will be needed in the final form of the test.

It is further assumed that tentative decisions have been reached as to $Pa_g$ and Pap, the probability of acceptance of good people, and the probability of acceptance of poor people. These decisions should be made in accordance with the principles described in the preceding chapter.

### Tentative Standardization of Test Items

Arrangements should be made to administer all items of the test to a group of people in order to determine item difficulty and item discrimination. The people chosen for this purpose should be a random selection from a population similar to those who will later take the test. If, for example, you plan to use the performance test at the end of the radar phase of the Class A school for electronics technicians, your standardization group should be a random selection of students in this course who have just finished the radar phase.

Each performance test item should be administered to each member of the standardization group. (It is desirable to administer all of the even numbered items, followed by all of the odd numbered items to half of the standardization group. The other half of the group would take the odd numbered items first, followed by the even numbered items. A test of significance of difference of mean scores should be computed

---

* The rationale for the empirical determination of item discrimination and item difficulties described in this chapter was developed independently by Dr. Lee J. Cronbach of the University of Illinois and by Dr. Jacob Wolfowitz of Columbia University. The mathematical formulae used here are those developed by Wolfowitz, based on the work of Dr. Abraham Wald. However, the description of the processes used is the responsibility of the present author, and any errors should be ascribed to him alone.

1. Adkins, Dorothy C., Construction and Analysis of Achievement Tests, 1947, Superintendent of Documents, Washington, D.C.
Micheels, W. J. and Karnes, M. R. Measuring Educational Achievement, 1950, McGraw Hill, New York.
U. S. Navy, Constructing and Using Achievement Tests, NAVPERS 16808. 1944, Bureau of Naval Personnel, Washington, D.C.

in order to determine whether it is tenable to assume that there is no appreciable amount of progressive learning occurring during the test.

a. If this hypothesis is untenable, the standardization scores should be based on the performance of only one-half of the standardization group, and in the future, test items should be administered in exactly the same order they were taken by the standardization group.
b. If this hypothesis is tenable, the magnitude of the difference in mean scores should be inspected. If the difference in mean scores is relatively large, and the N is small, you may wish to consider the null hypothesis untenable, even though this has not been demonstrated statistically, and hence administer the test items in standard order. If the N is reasonably large, and the difference in mean scores is relatively small, standardization scores should be based on the performance of the entire standardization group, and in the future, test items can be administered in any order.)

After the proposed performance items have been administered to the standardization group, the next step is to decide who are the "good" men and who are the "poor" men.

a. Total each man's raw score.
b. Arrange total scores in numerical order, with desirable scores first. Presumably, if the test is valid, the "good" men will be at the top, and "poor" men at the bottom.
c. Determine the score which separates the "good" men from those who are mediocre. This score is designated $m_g$. Determine the score which separates the really "poor" men from those who are mediocre. This score is designated $m_p$. (In most school situations, "good" men will be those who would score "A", "B", or "C", and really "poor" men would be those who would be failed. Mediocre men would be those who would receive a grade of "D".)
d. Put the names of the good men in one list, and names of the really poor men in a second list.

The last step in the standardization process is the determination of the discrimination scores for each item. The discrimination score may be abbreviated $D_s$, and is determined by dividing the proportion of poor people making a certain score on one item by the proportion of good people making the same score on that item.

a. Group the raw scores that it is possible to make on item number one, so that you have between five and ten groups.
b. Tabulate the number of good people who fall into each score group on item number one. Determine the proportion of good people in each group.
c. Tabulate the number of poor people who fall into each score group on item number one. Determine the proportion of poor people in each group.
d. For each raw score group on item number one, you should have two proportions. Divide the proportion found in "c" above by the proportion found in "b" above. (That is, divide the proportion of poor people in a certain score group on an item by the proportion of good people in that same score group on that

item.) The quotient is the discrimination score. This process converts each raw score into a discrimination score ($D_s$). These discrimination scores can range in value from zero to infinity.

e. Repeat steps "a" through "d" for each test item.

f. Check item number one to make sure that the discrimination scores form a sequence from high to low, with no reversals in value, and no $D_s$ values of infinity. If there are reversals or values of infinity, employ curve smoothing to eliminate them. Curve smoothing may be done by plotting the $D_s$ values and drawing a smooth curve by inspection, or by
   (1) averaging the proportions of good people in each set of three adjacent cells;
   (2) averaging the proportion of poor people in each set of three adjacent cells;
   (3) computing the $D_s$ values from these averages.

   Consider the following example:

Item Number One

| Raw Score | Good People Number | Good People Proportion | Poor People Number | Poor People Proportion | $D_s$ |
|---|---|---|---|---|---|
| 1-20 | 1 | .05 | 10 | .67 | 10.00 |
| 21-40 | 0 | .00 | 1 | .07 | infinity |
| 41-60 | 4 | .20 | 3 | .20 | 1.00 |
| 61-80 | 8 | .40 | 1 | .07 | .18 |
| 81-100 | 7 | .35 | 0 | .00 | .00 |
| total | 20 | 1.00 | 15 | 1.01 | |

There is a reversal between the $D_s$ corresponding to raw scores of 1-20 and raw scores of 41-60, since the $D_s$ values are not in numerical order. Moreover, there is one $D_s$ value of infinity. We can smooth these figures by recomputing the proportion of good people who made raw scores of 21-40 by averaging the original proportion, .00, with the two proportions on each side of it, (.05, corresponding to a score of 1-20; and .20 corresponding to a score of 41-60). This will yield an average proportion of .08. Repeat this for each of the proportions for both good and poor people using three adjacent proportions for each average. (For the highest and lowest scores, there are no data available for the third proportion. In this case, it is usually best to assume that the unknown proportion is the same as the last known proportion. Thus the smoothed proportion of good people corresponding to a raw score of 81-100 would be the average of .40, .35, and .35.)

If this procedure is followed, the example would appear like this:

Item Number One

| Raw Score | Good People Proportion | Poor People Proportion | $D_s$ |
|---|---|---|---|
| 1-20 | .03 | .47 | 15.7 |
| 21-40 | .08 | .31 | 3.88 |
| 41-60 | .20 | .11 | .55 |
| 61-80 | .32 | .09 | .28 |
| 81-100 | .37 | .02 | .05 |
| total | 1.00 | 1.01 | |

Occasionally, it may be desirable to go one step further, and employ curve smoothing on the $D_s$ values.
Reversals of the type described above are caused by too small a standardization sample, or by items which are unreliable.
Ideally, items which show reversals should be discarded, but in view of the small samples normally available for initial standardization, curve smoothing will usually give interpretable results. However, if the poor men make better scores than the good men, the item should be discarded, at least until more data can be obtained.

g. Repeat step "f" for each test item.
Discrimination scores obtained on the items which are retained after the original standardization process should not be regarded as fixed values, but should be corrected as additional data become available during the use of the test.

## Computation of A and B

In the previous chapter, considerable attention was paid to $Pa_g$ and $Pa_p$, the probability of acceptance of good and poor men, respectively. These values are used in computing A and B, which are the limits for discrimination scores, and determine when a person is accepted, rejected, or when additional testing needs to be done. These relationships are rather simple:

$$B = \frac{Pa_p}{Pa_g} = \text{Point of acceptance}$$

$$A = \frac{1 - Pa_p}{1 - Pa_g} = \text{Point of rejection}$$

TABLE 1

Values of A and B for Common Values of $Pa_g$ and $Pap$

| Probability of Acceptance of Good People $Pa_g$ | Probability of Acceptance of Poor People $Pa_p$ | A | B |
|---|---|---|---|
| .99 | .40 | 60.0 | 0.404 |
| | .30 | 70.0 | .303 |
| | .20 | 80.0 | .202 |
| | .10 | 90.0 | .101 |
| | .05 | 95.0 | .051 |
| .95 | .40 | 12.0 | 0.421 |
| | .30 | 14.0 | .316 |
| | .20 | 16.0 | .211 |
| | .10 | 18.0 | .105 |
| | .05 | 19.0 | .053 |
| .90 | .40 | 6.00 | 0.444 |
| | .30 | 7.00 | .333 |
| | .20 | 8.00 | .223 |
| | .10 | 9.00 | .111 |
| | .05 | 9.50 | .056 |
| .85 | .40 | 4.00 | 0.471 |
| | .30 | 4.667 | .353 |
| | .20 | 5.333 | .235 |
| | .10 | 6.00 | .118 |
| | .05 | 6.333 | .059 |
| .80 | .40 | 3.00 | 0.5 |
| | .30 | 3.5 | .375 |
| | .20 | 4.00 | .25 |
| | .10 | 4.5 | .125 |
| | .05 | 4.75 | .063 |
| .75 | .40 | 2.4 | .533 |
| | .30 | 2.8 | .4 |
| | .20 | 3.2 | .267 |
| | .10 | 3.6 | .133 |
| | .05 | 3.8 | .067 |

For example, if the probability of acceptance of good men were set at .95, and the probability of acceptance of poor men were set at .20, A and B could be determined as follows:

$$B = \frac{Pa_p}{Pa_g} = \frac{.20}{.95} = .21$$

$$A = \frac{1 - Pa_p}{1 - Pa_g} = \frac{1 - .20}{1 - .95} = \frac{.80}{.05} = 16.0$$

Values of A and B for a variety of common values of $Pa_g$ and $Pa_p$ are shown in Table 1.

## Scoring Performance Items During Routine Test Administration

As each man completes a performance test item, his raw score is determined, and then converted to a discrimination score, using a conversion table based on the standardization process described above. His discrimination score is then compared with the values determined for A and B. This will result in one of three actions:

1. If the man's discrimination score is equal to or greater than A, he is immediately rejected (flunked), and takes no more test items.

2. If the man's discrimination score is equal to or smaller than B, he is immediately accepted (passed), and takes no more test items.

3. If the man's discrimination score is between A or B, he proceeds to the second test item.

Suppose that for a particular man, action 3 is indicated. After he has completed test item number two, his raw score on this item is determined, and converted to a discrimination score. Since a man's score in a sequential test is based on all of the items he has taken previously during the test, we multiply the discrimination score he made on the second item by the discrimination score he made on the first item, and compare the result with A and B. This will again result in one of the three actions outlined above.

Suppose that after the second test item, action 3 is indicated again. Test item number three is administered, a discrimination score determined, and multiplied by the product of all previous discrimination scores ($D_s$ for item one X $D_s$ for item two, X $D_s$ for item three) and compared with A and B.

This process continues until the man is either accepted, or rejected, or until no more test items are available. If no more test items are available, the man is declared to be accepted, if there is a critical need for men; or rejected, if there is not a critical need for men.

## Example of Procedures In Sequential Analysis

Ten performance test items on trouble-shooting the 5G1b radar were prepared and administered to fifty-seven students who had just completed the radar section of a Navy Class A electronics technicians school. The following total scores were obtained:

| | | | | | |
|---|---|---|---|---|---|
| 148 | 96 | 74 | 62 | 56 | 36 |
| 132 | 93 | 74 | 62 | 56 | 34 |
| 130 | 90 | 74 | 61 | 55 | 34 |
| 123 | 89 | 73 | 60 | 54 | 30 |
| 115 | 88 | 72 | 59 | 54 | 29 |
| 114 | 88 | 70 | 59 | 51 | 27 |
| 110 | 87 | 69 | 59 | 47 | 25 |
| 104 | 84 | 66 | $m_g$ | 45 | 24 |
| 102 | 80 | 66 | 56 | 41 | |
| 98 | 76 | 65 | 56 | 36 $m_p$ | |

High scores indicated good performance.

It was decided arbitrarily that all men who scored above 58 were definitely good men, and that all those who scored below 37 were definitely poor men who needed additional training.

Discrimination scores were determined for item number one as follows:

### Item Number One

| Raw Score | Good People Number | Good People Proportion | Poor People Number | Poor People Proportion | $D_s = \frac{\text{Poor Prop.}}{\text{Good Prop.}}$ |
|---|---|---|---|---|---|
| 0 | 11 | .30 | 6 | .67 | 2.23 |
| 3 | 6 | .16 | 3 | .33 | 2.06 |
| 6 | 2 | .05 | 0 | .00 | 0 |
| 12 | 2 | .05 | 0 | .00 | 0 |
| 18 | 5 | .14 | 0 | .00 | 0 |
| 24 | 6 | .16 | 0 | .00 | 0 |
| 30 | 5 | .14 | 0 | .00 | 0 |
| total | 37 | 1.00 | 9 | 1.00 | |

Similar procedures were followed for the remaining nine items, and the following discrimination scores were obtained:

Item No. Two

| Raw Score | $D_s$ |
|---|---|
| 0 | 1.57 |
| 1-6 | 1.37 |
| 8-10 | .29 |

Item No. Three

| Raw Score | $D_s$ |
|---|---|
| 0 | 2.20 |
| 1-6 | 2.05 |
| 8-10 | .19 |

Item No. Four

| Raw Score | $D_s$ |
|---|---|
| 0 | 4.40 |
| 2 | 1.57 |
| 4-6 | .94 |
| 8 | .92 |
| 10 | 0 |

Item No. Five

| Raw Score | $D_s$ |
|---|---|
| 0-3 | 2.17 |
| 5-12 | .79 |
| 18-30 | 0 |

Item No. Six

| Raw Score | $D_s$ |
|---|---|
| 0 | 2.97 |
| 4-8 | .58 |
| 12-20 | 0 |

Item No. Seven

| Raw Score | $D_s$ |
|---|---|
| 0 | 2.75 |
| 4-8 | 1.38 |
| 12-20 | .37 |

Item No. Eight

| Raw Score | $D_s$ |
|---|---|
| 0 | 2.33 |
| 2-4 | 1.38 |
| 8 | 1.00 |
| 12-20 | 0 |

Item No. Nine

| Raw Score | $D_s$ |
|---|---|
| 0-2 | 2.48 |
| 4 | 2.20 |
| 8-12 | .69 |
| 16-20 | 0 |

Item No. Ten

| Raw Score | $D_s$ |
|---|---|
| 0-3 | 1.59 |
| 6 | 1.38 |
| 12-18 | 1.00 |
| 24-30 | 0 |

Since there were only nine "poor" people in the sample, these discrimination values are regarded as only tentative. They should be corrected as additional data are available.

## Application of $D_s$ Values

If it were decided to use these particular performance test items in some future testing program the procedure would be as follows:

a. Determine $Pa_g$ and $Pa_p$ Suppose that since the Navy needed all of the good ET's it could get, a decision was made to risk failing only 5 per cent of the good men. Thus the probability of acceptance for good men ($Pa_g$) would be .95. Since there would be further opportunities for screening men at later dates, it might be decided to take a risk of accepting 20 per cent of the poor men. Thus the probability of accepting poor men would be .20.

Referring to Table I, these figures give an "A" of 16.0, and a "B" of .211.

b. Give one test item to the first man tested. The first man tested could be given any one of the ten items. (Suppose that item number five were used first, and the man made a raw score of two on it.)

c. Determine the $D_s$ value of the raw score made on the item just taken. (The $D_s$ value of a raw score of two on item number five from the preceding tables is 2.17.)

d. Compare this $D_s$ value with the A and B determined in step "a" above. Stop testing and reject the man if the $D_s$ value is equal to or greater than A. Stop testing and accept him if the $D_s$ value is equal to or smaller than B.

(Since 2.17 is between 16 and .211, we would continue testing our man.)

e. If the decision is to continue testing, administer a second item, and repeat step "c" above. (Suppose that item number three was administered, and a raw score of five obtained. This would give a $D_s$ value of 2.05.)

f. Multiply the $D_s$ values for the first and second items taken, and repeat step "d" above. (2.05 times 2.17 is 4.85, so the decision is to continue testing.)

g. If additional items are needed, administer them one at a time. Multiply the $D_s$ value for the latest item taken by the result of all previous multiplications of $D_s$ values. After each item, repeat step "d" above. (Suppose that the third item administered was item number four, and a raw score of zero was obtained. This has a $D_s$ value of 4.40. 4.40 times 4.85 (obtained in step "f") is 21.34, so the man is rejected and testing stopped.)

## Suggested Modification of Scoring Performance Items During Routine Test Administration

Cronbach has suggested that fewer errors are apt to result during routine test administration if logarithms of A, B, and $D_s$ values are used. If this procedure is employed, $D_s$ values can be added instead of being multiplied. A simple experiment with personnel of the type who will administer the performance test should quickly indicate whether errors of multiplication or errors of addition (using positive and negative numbers) are most important.

## Estimating an Operating Characteristics Curve

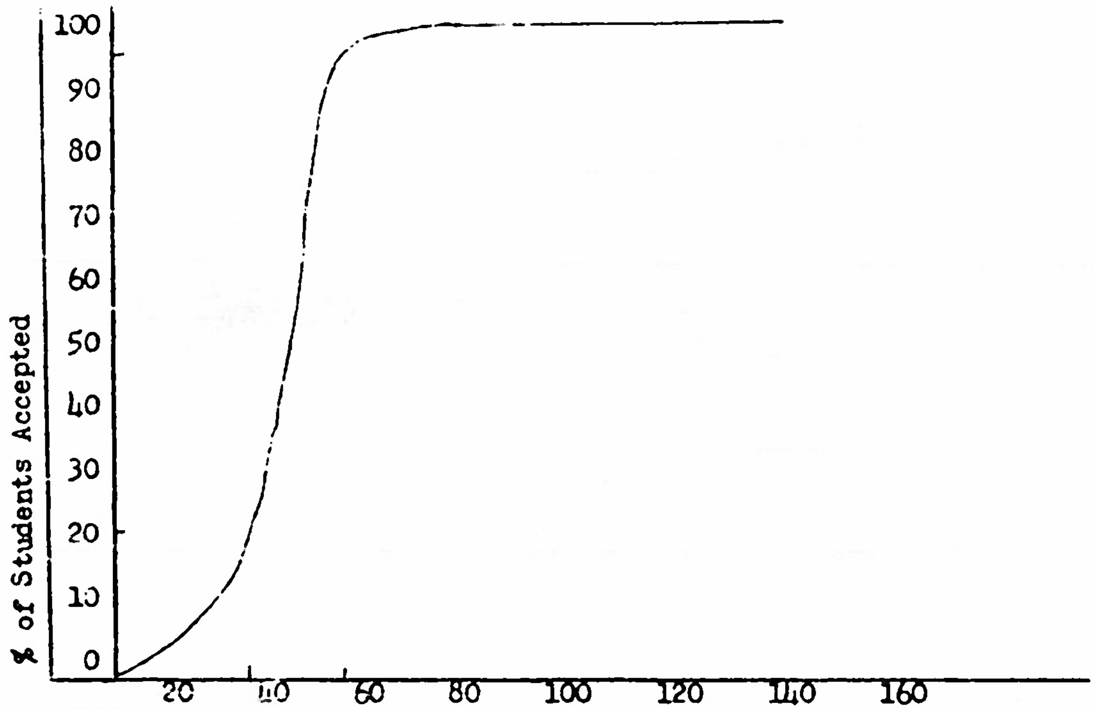
The operating characteristics curve is a graphic representation of the efficiency of any test. Its use was described in Chapter II.

For most practical purposes, the operating characteristics curve for a sequential sampling plan can be determined from four points. Two of these, Pag - $P_g$ and Pap - Pp, have been determined previously. The other two are established by the facts that people who have zero true ability will never be accepted, and that people who have perfect true ability will always be accepted.

If $P_{ag}$ = .95, $m_g$ = 58; and $P_{ap}$ = .20, $m_p$ = 37, the operating characteristics curve would appear as follows:

Operating Characteristics Curve
For Sequential Sampling Plan Determined By

$P_{ag}$ = .95, $m_g$ = 58, $P_{ap}$ = .20, $m_p$ = 37




Students' True Score

If additional points on the oc curve are needed for increased accuracy, they may be obtained through a process outlined in Sequential Analysis of Statistical Data: Applications, 1945, Columbia University Press, pages 4.19 -4.23.

This process consumes a considerable amount of time. It is usually unnecessary to graph even the simplified oc curve shown here except when the effects of choosing different $m_g$, $m_p$, $P_{ag}$ and $P_{ap}$ values are to be compared.

Computation of Average Sample Size Required

The complete average sample size curve for a given sequential sampling plan can be determined from formulae described on pages 4.20-4.23 of the Columbia University publication described above. For ordinary purposes, however, the average sample size can be determined empirically from data provided by the standardization group.

The procedure for this determination is as follows:

a. Complete the administration of the test to the standardization group, and calculate A and B from your determination of $Pa_p$ and $Pa_g$.
b. Consider one man at the time from the standardization group. Multiply accumulatively the $D_s$ value he obtains on each test item. Starting with item number one, go through each of the items he took, and determine the item on which he was first failed or first accepted. If this was on the fifth item he took, record the number five.
c. Do the same for each man, and determine the average number of items it took to reach a decision.

You should be prepared to administer about three times the average number of items required by the standardization group. If this value is too large, lower A or raise B.

For example, if John Doe, in your standardization group, had the following scores:

| Sequence in Which Items Were Taken | $D_s$ Value | Cumulative Multiplication of $D_s$ Values |
|---|---|---|
| 1 | 2.1 | 2.1 |
| 2 | .7 | 1.47 |
| 3 | 3.4 | 5.14 |
| 4 | 2.7 | 13.88 |
| 5 | 4.0 | 55.52 |
| 6 | 4.0 | 222.1 |
| 7 | .3 | 66.6 |
| 8 | 2.0 | 133.2 |
| 9 | 1.7 | 226.4 |
| 10 | 3.4 | 769.8 |
| 11 | 1.7 | 1308.7 |
| 12 | 1.8 | 2355.7 |

If A were 20, and B were .05, this person would be rejected on the fifth item.

Note that the average sample size will increase if A and B are farther apart, and will decrease if A and B are closer together.

The above procedure will be satisfactory if items are administered in the same order as was used in the standardization procedure. It will probably be satisfactory, even if the order of administration is changed, provided that each of the items has approximately the same discrimination value.

## Alternative Procedure for Calculating Average Sample Size *

A considerably more accurate, but slightly more involved procedure would involve the computation of the geometric mean of $D_s$ values for each man, and a determination of the power to which this mean would have to be raised in order to approximate the A value used for the poor group, and the B value used for the good group. (This procedure is not applicable if any of the smoothed $D_s$ values are zero or infinity.)

a. Complete the administration of the test to the standardization group, and calculate A and B from your determination of $Pa_p$ and $Pa_g$.
b. Determine log A and log B.
c. Determine the log of each $D_s$ value in your smoothed standardization data.
d. Calculate the log of the geometric mean of $D_s$ values earned by the poor group.

$$\text{Log geometric mean of } D_s \text{ values for poor group} = \frac{\sum \sum \text{Log } D_s \text{ poor group}}{\text{No. of men in poor group} \times \text{No. of items per man}}$$

e. Calculate the log of the geometric mean of $D_s$ values earned by the good group.

$$\text{Log geometric mean of } D_s \text{ values for good group} = \frac{\sum \sum \text{Log } D_s \text{ good group}}{\text{No. of men in good group} \times \text{No. of items per man}}$$

f. Divide log A by the value obtained in "d" above. This is the average number of items required to fail a poor man.
g. Divide log B by the value obtained in "e" above. This is the average number of items required to accept a good man.
h. Multiply the value obtained in "f" above, by the proportion of poor men. (If you have 10 poor man and 40 good men, the proportion of poor men is .20; disregard the "indifferent" men.)
i. Multiply the value obtained in "g" above by the proportion of good men.
j. Add the values obtained in "h" and "i" above. You should be prepared to administer approximately three times this number of test items to some men. If this value is too large, lower A or raise B.

---

* This procedure, based on information theory, is suggested by Cronbach.

## Determination of Minimum Number of Items Required to Pass or Fail a Testee

To determine the minimum number of items necessary to fail a student, arrange the five or six highest $D_s$ values in rank order. If the largest $D_s$ value is larger than A, a person can be failed after taking only one item. If the largest $D_s$ value is smaller than A, multiply it by the second largest, and again compare with A. Continue until a value as large or larger than A is obtained. The minimum number of items necessary to fail a student is equal to the number of $D_s$ values multiplied together to exceed A.

To determine the minimum number of items necessary to pass a student, arrange the five or six lowest $D_s$ values in rank order. If the smallest $D_s$ value is smaller than B, a person can be passed after taking only one item. If the smallest $D_s$ value is larger than B, multiply it by the next smallest, and again compare with B. Continue until a value as large or larger than B is obtained. The minimum number of items required to fail a student is equal to the number of $D_s$ values multiplied together to reach a value less than B.

## Summary

Sequential sampling appears to be useful in testing, whenever:

1. Testing time per test item is high in relation to the time required to score each test item, and
2. The test is primarily designed to determine whether a person "passes" or "fails", and
3. There is a need for testing more than about one hundred persons on the same test, either in one group or in a number of groups, and
4. There is negligible correlation between items when criterion scores are held constant.

Sequential sampling takes item difficulty and item discrimination into account when discrimination score values (norms) are established.

Sequential sampling can readily be adapted to changing standards of accepting people, with no revision of the norms previously set up.

Ordinarily, sequential sampling will give about the same accuracy as a fixed length test, with about half of the testing time, and about half of the testing cost.

# APPENDIX

## Results of Sequential Sampling Compared with Administration of a Fixed Length Test

As a check on the efficiency of sequential sampling, the Ds values obtained from the standardization process described in the last chapter were applied in sequential fashion to the scores obtained. The following results were obtained, using A = 16, B = .211:

| MAN NO. | TOTAL RAW SCORE | CLASSIFICATION ON TOTAL RAW SCORE | CLASSIFICATION ON SEQUENTIAL SCORE | NO. OF SEQUENTIAL ITEMS REQUIRED |
|---|---|---|---|---|
| 47 | 148 | good | pass | 1 |
| 46 | 132 | " | " | 1 |
| 24 | 130 | " | " | 5 |
| 19 | 123 | " | " | 4 |
| 43 | 115 | " | " | 1 |
| 36 | 114 | " | " | 7 |
| 10 | 110 | " | " | 2 |
| 27 | 104 | " | " | 1 |
| 50 | 102 | " | " | 2 |
| 35 | 98 | " | " | 5 |
| 4 | 96 | " | " | 4 |
| 8 | 93 | " | " | 3 |
| 5 | 90 | " | " | 7 |
| 41 | 89 | " | " | 4 |
| 51 | 88 | " | " | 6 |
| 14 | 88 | " | " | 7 |
| 48 | 87 | " | " | 4 |
| 32 | 84 | " | " | 5 |
| 23 | 80 | " | " | 7 |
| 6 | 76 | " | " | 7 |
| 16 | 74 | " | " | 4 |
| 17 | 74 | " | " | 5 |
| 44 | 74 | " | " | 1 |
| 39 | 73 | " | " | 4 |
| 20 | 72 | " | " | 10 |
| 21 | 70 | " | " | 5 |
| 29 | 69 | " | fail | 7 |
| 49 | 66 | " | pass | 3 |
| 11 | 66 | " | " | 3 |
| 55 | 65 | " | " | 9 |
| 38 | 62 | " | fail | 3 |
| 15 | 62 | " | pass | 10 |
| 34 | 61 | " | " | 2 |
| 57 | 60 | " | " | 2 |
| 1 | 59 | " | " | 2 |
| 26 | 59 | " | fail | 5 |

| MAN NO. | TOTAL RAW SCORE | CLASSIFICATION ON TOTAL RAW SCORE | CLASSIFICATION ON SEQUENTIAL SCORE | NO. OF SEQUENTIAL ITEMS REQUIRED |
|---|---|---|---|---|
| 42 | 59 | good | pass | 5 |
| 7 | 56 | indifferent | " | 7 |
| 3 | 56 | " | " | 10 |
| 40 | 56 | " | fail | 7 |
| 37 | 56 | " | pass | 3 |
| 22 | 55 | " | fail | 6 |
| 18 | 54 | " | pass | 8 |
| 52 | 54 | " | fail | 6 |
| 13 | 51 | " | pass | 1 |
| 30 | 47 | " | " | 5 |
| 2 | 45 | " | " | 8 |
| 25 | 41 | " | fail | 9 |
| 54 | 36 | poor | " | 4 |
| 56 | 36 | " | " | 7 |
| 9 | 34 | " | " | 4 |
| 12 | 34 | " | " | 8 |
| 31 | 30 | " | " | 8 |
| 33 | 29 | " | " | 7 |
| 53 | 27 | " | " | 6 |
| 45 | 25 | " | " | 3 |
| 28 | 24 | " | " | 5 |

The average number of items required to reach a decision in sequential sampling was only 4.42 instead of 10, a saving of over three hours in average performance testing time. The biserial correlation between original total raw score and sequential pass-fail was .83. However, it should be noted that this correlation is somewhat contaminated, and in order to be verified, should be re-computed on data not used for standardization of the test.